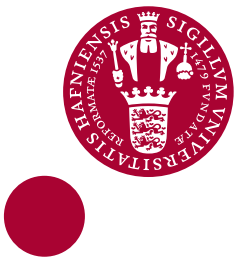


Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data

Waples, Ryan K.; Albrechtsen, Anders; Moltke, Ida

Published in:
Molecular Ecology

DOI:
10.1111/mec.14954

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Received: 21 March 2018 | Accepted: 12 October 2018

DOI: 10.1111/mec.14954

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data

Ryan K. Waples | Anders Albrechtsen | Ida Moltke

Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen N, Denmark

Correspondence
Anders Albrechtsen and Ida Moltke, Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen N, Denmark.
Emails: albrecht@binf.ku.dk; ida@binf.ku.dk

Funding information
RKW and IM were supported by a grant from the Independent Research Fund Denmark awarded to IM (DFF – 4090-00244). AA was supported by a grant from the Lundbeck Foundation (R215-2015-4174).

Abstract

Knowledge of how individuals are related is important in many areas of research, and numerous methods for inferring pairwise relatedness from genetic data have been developed. However, the majority of these methods were not developed for situations where data are limited. Specifically, most methods rely on the availability of population allele frequencies, the relative genomic position of variants and accurate genotype data. But in studies of non-model organisms or ancient samples, such data are not always available. Motivated by this, we present a new method for pairwise relatedness inference, which requires neither allele frequency information nor information on genomic position. Furthermore, it can be applied not only to accurate genotype data but also to low-depth sequencing data from which genotypes cannot be accurately called. We evaluate it using data from a range of human populations and show that it can be used to infer close familial relationships with a similar accuracy as a widely used method that relies on population allele frequencies. Additionally, we show that our method is robust to SNP ascertainment and applicable to low-depth sequencing data generated using different strategies, including resequencing and RADseq, which is important for application to a diverse range of populations and species.

KEYWORDS
ascertainment bias, IBD, identity by descent, low-depth, NGS, non-model, relatedness

1 | INTRODUCTION

The ability to infer the familial relationship between a pair of individuals from genetic data plays a key role in several research fields. In conservation biology, it is used to design breeding programmes that minimize inbreeding (Kardos, Luikart, & Allendorf, 2015), in archaeology it is helpful to understand burial patterns and other cultural traditions (Baca, Doan, Sobczyk, Stankovic, & Weglenski, 2012; Sikora et al., 2017), and in population and disease genetics it is often used to exclude relatives, because many analysis methods within those fields assume all analysed individuals are unrelated and violations of this assumption can lead to wrong conclusions (Balding, 2006).

Numerous pairwise relatedness inference methods have been developed, for example, Thompson (1975), Lee (2003), Purcell et al. (2007), Albrechtsen et al. (2009), Manichaikul et al. (2010), Stevens et al. (2011), Korneliussen and Moltke (2015), Conomos, Reiner, Weir, and Thornton (2016), Dou et al. (2017), and many are available in popular software packages, like PLINK (Purcell et al., 2007), and KING (Manichaikul et al., 2010). Most of these methods estimate either the three relatedness coefficients k_0, k_1 and k_2, or the kinship coefficient $\theta = \frac{k_1}{4} + \frac{k_2}{2}$ for each pair of diploid individuals, where k_0, k_1 and k_2 are the proportions of the genome where a pair of individuals share 0, 1 or 2 alleles identical by descent (IBD) (Thompson, 2000). By definition, alleles are IBD when they

are identical due to *recent* common ancestry, but because alleles can also be identical due to older common ancestry or recurrent mutations, IBD status cannot be directly observed. Therefore, the pairwise relatedness coefficients and the kinship coefficient have to be estimated, which can be done from patterns of observed genetic identity (identity by state; in short IBS). Once estimated, these relatedness statistics, k_0, k_1, k_2 and the kinship coefficient can be used to infer familial relationships by comparison with the expectation of the statistics for different familial relationships (Hill & Weir, 2011).

Although inference of relatedness is of wide interest, most existing methods are not immediately applicable in studies with limited data or genetic resources. First, most existing methods require the allele frequencies of the source population. For most studies in modern humans, this is not a problem. However, in studies of ancient humans or other species, accurate estimates of population allele frequencies are often not obtainable, because only a low number of samples are available. Second, several existing methods consider consecutive loci jointly and use sliding windows or hidden Markov models to leverage the non-independence of allele sharing along the genome between relatives (Albrechtsen et al., 2009; Gusev et al., 2009; Kuhn, Jakobsson, & Gunther, 2018; Stevens et al., 2011). These methods are powerful, but require information about the genomic position of variable sites, and for non-model organisms, high-quality reference genome assemblies are often not available. Third, other methods avoid allele frequencies, but rely on access to many samples to provide a necessary context for relationship classification (e.g., Abecasis, Cherny, Cookson, & Cardon, 2001). Finally, nearly all existing methods—both frequency-based and others—require genotype data. However, in sequencing studies, samples are often only sequenced to low depth due to cost and technical issues. This makes it infeasible to call genotypes accurately (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012), precluding the use of these methods. There are a few methods that estimate relatedness from low-depth sequencing data by utilizing genotype likelihoods (e.g., Korneliussen & Moltke, 2015), or by using imputed genotype dosages (Dou et al., 2017). However, these methods function by leveraging access to many samples to estimate allele frequencies or perform accurate genotype imputation and are therefore not designed to apply to data sets with a low number of samples.

Hence, most existing methods to infer close familial relationships are not immediately applicable in studies where data are limited, including many studies of non-model organisms and ancient samples. One of the few exceptions is a simple but elegant test for pairwise relatedness proposed in Lee (2003). This test relies entirely on the relative frequency of different genotype combinations within a pair of individuals and thus only requires genotype data from the two target individuals. While useful, this test does not provide any means to distinguish between different types of close familial relationships; it only provides a statistical test for a pair of individuals of the null hypothesis of them being unrelated.

There are only a few methods that can be used to distinguish between different types of familial relationships for a pair of individuals when neither allele frequencies nor information about the relative genomic position of sites is obtainable. One such method consists of plotting the proportion of the genomic sites in which the two individuals share both alleles IBS (which we will denote IBS2) vs. the proportion of the genomic sites in which they share zero alleles IBS (which we will denote IBS0). This method was used in Rosenberg (2006), where it was applied to the Human Genome Diversity Project (HGDP) data set. In the resulting scatter plot of the HGDP data (Figure 1 of Rosenberg (2006)), pairs of individuals with the same relationship category form distinct clusters so that it is possible to locate parent–offspring pairs, full-sibling pairs and to a lesser extent more distant relationships such as half-siblings/avuncular/grandparent–grandchild and first cousins.

Another such method is based in part on the KING-robust kinship estimator (Manichaikul et al., 2010). The KING-robust kinship estimator was developed to be robust to population structure, but in practice it has been shown to provide biased kinship estimates when applied to pairs of samples whose four chromosomes are not all from the same population (Conomos et al., 2016; Thornton et al., 2012). However, the KING-robust kinship estimator is directly applicable to samples from the same homogenous population even when allele frequencies are unknown. The reason for this is that, like the test suggested in Lee (2003), it relies only on the genotype combinations within the two target individuals and does not require knowledge about of allele frequencies. Importantly, Manichaikul et al. (2010) show it is possible to infer if a pair of individuals are parent–offspring, full-sibling, half-siblings/avuncular/grandparent–grandchild, first cousins, or unrelated, by jointly considering KING-robust kinship and the fraction of sites IBS0 using SNP array data without allele frequencies. For example, see Figure 3a in Manichaikul et al. (2010); a scatterplot of KING-robust kinship vs. the fraction of sites IBS0.

However, both the methods described above have two important limitations. First, like most other methods to estimate relatedness, they were developed for genotype data only. For example, the KING software (Manichaikul et al., 2010) implementing the KING-robust kinship estimator requires genotype data as input, which can be problematic for studies where only moderate or low-depth sequencing data are available and calling genotypes is consequently difficult. Second, both methods rely on estimates of the fraction of sites IBS0, which can be problematic because this fraction, as well as the fraction of sites IBS1 and IBS2, is highly sensitive to SNP ascertainment. This means that the results of the methods are platform-dependent and are likely to differ between different SNP arrays and especially between SNP array and sequencing data sets. In turn, this means that it can be difficult to distinguish between full-siblings and parent–offspring pairs using these methods.

Motivated by the outlined limitations to the existing methods, we present a method for relationship inference that–unlike most existing methods–relies neither on allele frequencies nor on information about the relative position of the variant sites, and which–unlike other frequency-free methods— is (1) applicable even to sequencing data of so low depth that accurate genotypes cannot be called from it and (2) robust to SNP ascertainment bias.

The new method is inspired by previous methods; it uses the KING-robust kinship estimator and a statistic R0, which is similar to the test statistic from the test for relatedness suggested by Lee (2003). However, the method is new in two important ways. First, besides relying on the two statistics, R0 and KING-robust kinship, it also relies on a third new statistic, R1. More specifically, the method consists of using two combinations of these three statistics, R1–R0 and R1-KING-robust kinship, to infer relationships, and it is this combination of statistics that makes the method robust to ascertainment bias. Second, while the new method is straightforward to apply to genotype data like other similar methods, we also present two computational approaches to estimate the three statistics directly from sequencing data that take the uncertainty of genotypes into account, allowing application to low-depth sequencing data.

In the following, we first fully describe the three statistics, R0, R1 and KING-robust kinship, how they can be estimated and other methodological details. Next, using simulated and publicly available SNP array data, we show that the new method provides similar accuracy and precision to the commonly used frequency-based method implemented in PLINK, when such data are available. Then, using sequence data from the 1,000 Genomes Project (The Genomes Project, 2015), we show that the three statistics can be estimated directly from sequencing data of low depth (~4×), here defined as depth insufficient for accurate genotype calling. Moreover, we show that the estimates obtained in this way are useful for inference of close familial relationships and that this is not the case for estimates obtained from genotypes called from the same data. Using different subsets of the same data, we also show that this new method, unlike previous similar methods, is robust to SNP ascertainment. Finally, we show that the method also provides useful results when applied to sequencing data down-sampled to approximate data generated using reduced-representation approaches, for example, restriction site-associated DNA sequencing (RADseq) and discuss some potential applications and limitations of the new method.

2 | METHODS AND MATERIALS

2.1 | The R0, R1 and KING-robust kinship statistics

The method for relationship inference we propose consists of estimating three statistics called R0, R1 and KING-robust kinship from genetic data and interpreting plots of R1 vs. R0 and R1 vs. KING-robust kinship.

We define the three statistics, R0, R1 and KING-robust kinship in terms of the genomewide IBS-sharing pattern of two individuals of interest. At any given diallelic site, a pair of individuals will carry one of nine possible genotype combinations; the nine possible combinations of the two individuals each carrying 0, 1 or 2 copies of a specific allele, for example, the ancestral allele. We can therefore fully characterize the genomewide IBS-sharing pattern of a pair of individuals by nine counts or proportions denoted: A, B, C, D, E, F, G, H and I (Figure 1a), similar to a two-dimensional site-frequency spectrum (SFS) across the two individuals. The R0 and R1 statistics are defined as simple functions of a subset of these nine values as shown in Figure 1b,c, and the KING-robust kinship statistic, originally defined by Manichaikul et al. (2010), can also be re-formulated as a function of these 9 values (Figure 1d).

The new method is motivated by several observations. First, the expected values of A–I vary depending on the familial relationship between the pair of individuals of interest. Consequently, so do functions of A–I, including R0, R1 and KING-robust kinship. Notably, there is no overlap between the joint expectation ranges of [R1, R0] and [R1, KING-robust kinship] for the four close relationship categories: full-siblings (FS), half-siblings/avuncular/grandparent–grandchild (HS), first cousins (C1) and unrelated (UR) and the range of expected values for parent–offspring (PO) only overlaps with those of FS in a single point (Figure 2, for derivations see supplementary text). Crucially, this is true regardless of the underlying allele frequency spectrum and holds for any pair of non-inbred individuals from the same homogenous population, making [R1, R0] and [R1, KING-robust kinship] potentially useful for distinguishing between these relationships. Second, while A–I, and thus R0, R1 and the KING-robust kinship estimator can be calculated from genotype data, they can also be estimated directly from next-generation sequencing (NGS) data based on the expected number of sites with each genotype combination (see below for details). This makes the method appropriate even when the sequencing depth is too low for accurate genotype calling (see below for methodological details). Third, regardless of the type of data that is available, R0, R1 and KING-robust kinship can be estimated without the need for population allele frequencies or information about the relative position of the genomics sites analysed. Finally, we expect the three statistics to be robust to SNP ascertainment because they are ratios computed from sites that are variable within the two samples and should thus be unaffected by the number of non-variable sites and because the (unknown) underlying frequency spectrum should only have a limited effect on these ratios.

2.1.1 | Estimation from sequencing data

The counts of the nine genotype combinations, A–I, and thus R0, R1 and KING-robust kinship for a pair of individual, can be estimated directly from NGS data via the use of genotype likelihoods calculated from aligned sequencing reads. Genotype likelihoods provide a means to account for the genotype uncertainly inherent to low-depth NGS data. We used two distinct, but similar, approaches to estimate these statistics from sequencing data that both build on this idea.

The first approach, which we denote the IBS-based approach, considers all ten possible genotypes at each diallelic site for each of the two individuals of interest and consists of a maximum-likelihood (ML) estimation of the counts of each of the 100 (10 × 10) possible genotype pairs (for details, see supplementary text). To perform the ML estimation, we used an expectation–maximization (EM) algorithm, which we have added to the ANGSD software package (Korneliussen et al., 2014) "IBS". After obtaining the estimate of the

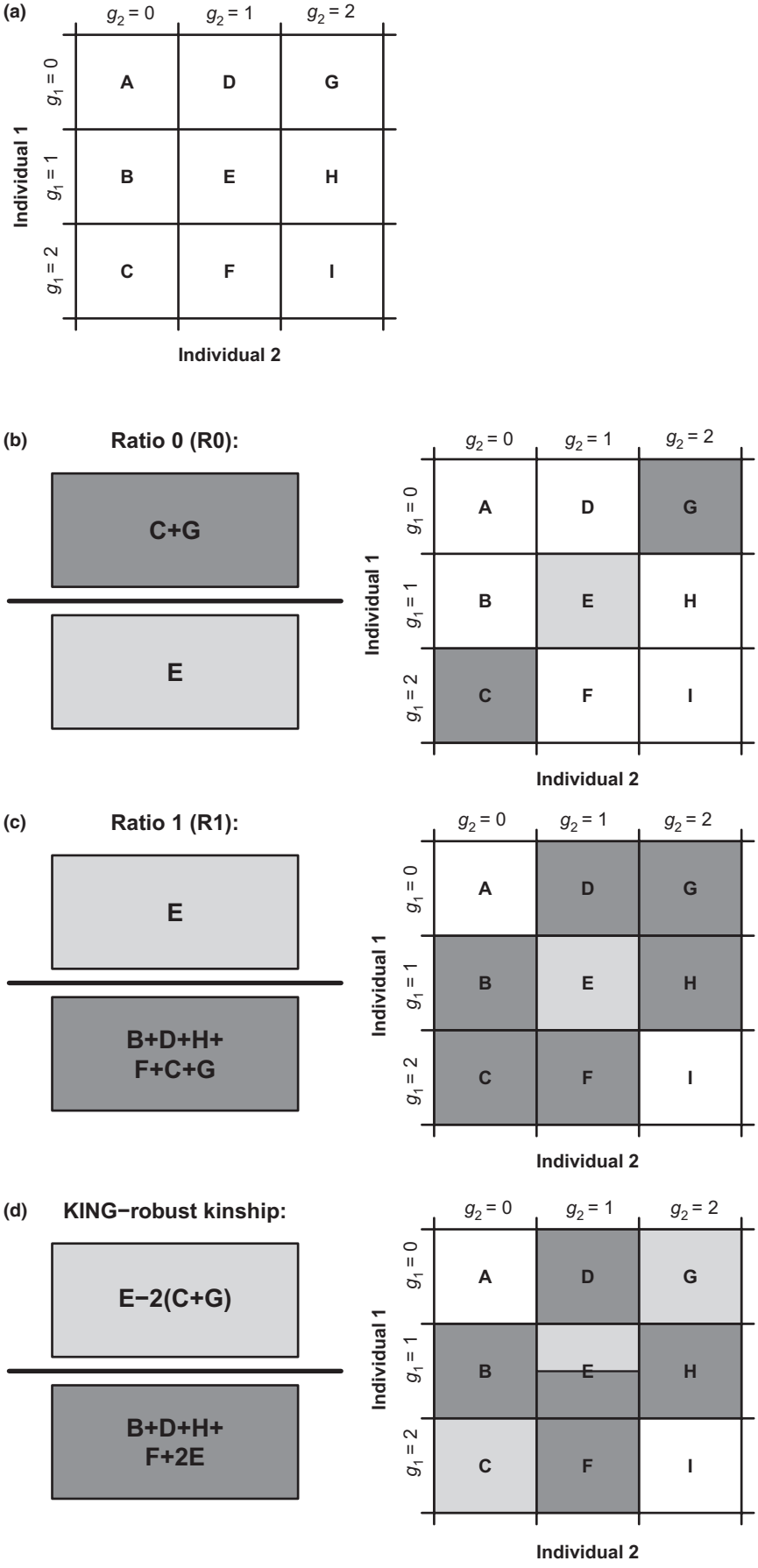


FIGURE 1 Definitions of pairwise genotype categories A–I and the R0, R1 and KING-robust kinship statistics. (a) Definition of the pairwise genotype categories A–I. Here, g_1 and g_2 denote the numbers of genotype for each of the two diploid individuals, 1 and 2, respectively. These genotypes are defined as the number of copies of a certain allele carried by individual 1 and 2, respectively. We assume diallelic variants such that g_1 and g_2 each has 3 possible values: 0, 1 and 2. For a pair of individuals, there are nine possible genotype combinations. We organize them into a 3 × 3 matrix and denote them with the letters from A to I. The values A–I can equivalently be either counts or proportions. (b) Definition of the R0 statistic based on the notation illustrated in (a). (c) Definition of the R1 statistic based on the notation illustrated in (a). (d) Definition of the KING-robust kinship estimator (Manichaikul et al., 2010), formulated using the notation illustrated in (a)

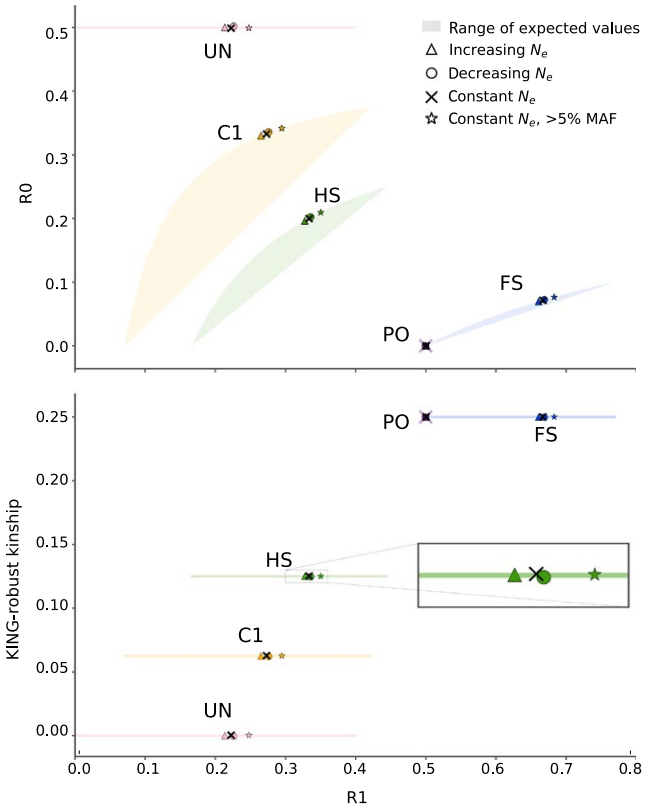


FIGURE 2 Ranges of expected values and simulation results for R1–R0 and for R1-KING-robust kinship for each of five relationship categories: parent–offspring (PO), full-siblings (FS), half-siblings/avuncular/grandparent–grandchild (HS), first cousins (C1) and unrelated (UR). (Top) The coloured shaded areas (sometimes just lines) show theoretically derived ranges of the joint expectation for R1–R0 based on expected IBD sharing (i.e., values of k_0, k_1 and k_2) for each relationship across all possible allele frequency spectra. For PO, this range is a singular point and is shown as a shaded purple X. The coloured symbols (triangle, circle, star) and black "X"s show values for each relationship obtained from data simulated under four different scenarios. Three of the scenarios are different demographic histories: (1) a 10-fold increase in *Ne* over the last 100 generations, (2) constant N_e and (3) a 10-fold decrease in N_e over the last 100 generations. The fourth scenario is also constant N_e, but sites are ascertained to have allele frequency above 5%. Note that, while it is difficult to see due to overplotting, all simulated values for PO fall very close to (R1, R0) = (0.5, 0). (Bottom) same as (Top), but for (R1, KING-robust kinship). Here all simulated values for PO fall very close to (R1, KING-robust kinship) = (0.5, 0.25)

counts of all the 100 possible genotype pairs, we converted them into estimates of A–I, by summing over the counts that correspond to each combination. For example, the genotype pairs: AA/AA, CC/CC, GG/GG and TT/TT all contribute to cells A or I of Figure 1. Counts corresponding to genotype pairs with more than two different alleles (e.g., AC/AG) were discarded. The advantage of this IBS-based approach is that it does not require specification of a known allele at each site and can thus be applied to nearly any sequencing data set, even with low-depth, without any prior information on the alleles at each site.

The second approach, which we denote the SFS-based approach, consists of performing ML estimation of the two-dimensional site-frequency spectrum (SFS) as in realSFS (Nielsen et al., 2012). To find the ML estimate of the SFS, we used an EM method implemented in the ANGSD software package under the name "realSFS" (Korneliussen, Albrechtsen, & Nielsen, 2014). The SFS-based approach requires one allele to be specified for each site, for example, the ancestral, the consensus or the reference allele. The model underlying this approach assumes that genotypes for each site have the possibility of containing this specified allele and up to one other unspecified allele. For the analyses performed in this paper, we used the consensus sequences from the highest depth individual (NA19042) to specify the alleles that exist at each site and restricted our analysis to sites where the depth in this individual was at least three.

The computational burden of analysing genomewide data sets can be significant. For both the genotype likelihood-based approaches described above, the main limitation is RAM, as data likelihoods for each site need to be loaded into memory for optimization. To overcome this limitation, we analysed each chromosome separately and then combined the values for each chromosome to produce a genomewide estimate. To calculate genotype likelihoods of the sequencing data, we used the original GATK genotype likelihood model (Mckenna et al., 2010) with independent errors, as implemented in ANGSD. We also tried to use the samtools genotype likelihood model (Li, 2011) with a more complicated error structure, but found it produced worse results (data not shown). Both IBS and realSFS produce the expected values for A–I, which we subsequently use to calculate R0, R1 and KING-robust kinship. For example command lines for each analysis, see supplemental text.

2.1.2 | Confidence intervals

All of the above estimation methods treat each site as independent. This assumption should not affect our expectation of each statistic (Wiuf, 2006), but statistical non-independence (here due to linkage disequilibrium (LD) and IBD) does affect standard estimates of uncertainty. To quantify the uncertainty, we therefore estimated confidence intervals for all statistics using a block-jackknife procedure. Confidence intervals were estimated by leaving each chromosome out (chromosome jackknife), which takes both the IBD and the LD correlation into account. The weighted block-jackknife variance estimator of Busing, Meijer, and Leeden (1999) was used to estimate the variance from the distribution of estimates for each statistic. The square root of this variance was interpreted as the standard error in our estimate.

2.2 | Application to simulated data

To evaluate the new method and to investigate the effects of demography on the expected statistics, we simulated genotype data under three different demographic histories: (1) a demography with constant effective population size (N_e), (2) a shrinking demography

with a 10-fold decrease in N_e over the last 100 generations, (3) an expanding demography with a 10-fold increase in N_e over the last 100 generations. For each of the three scenarios, we used the coalescent simulator msprime (Kelleher, Etheridge, & McVean, 2016) to simulate four haplotypes. These haplotypes were then used to construct pairs of related individuals in five relationship categories: parent–offspring (PO), full-siblings (FS), half-siblings/avuncular/grandparent–grandchild (HS), first cousin (C1) and unrelated (UR). For unrelated pairs of individuals, the genotype data were constructed by simply splitting the four haplotypes into two pairs. For related pairs of individuals, we constructed the genotypes at each variable site by first sampling whether the individuals shared 0, 1 or 2 alleles IBD according to the expected values of $[k_0, k_1, k_2]$ for the relevant relationship (Supplementary Table S1). Then, we used alleles present on the four haplotypes according to the sampled IBD status to construct the genotypes at that site. For example, if the individuals were sampled to share two alleles IBD at a site, both individuals were assigned the genotype consisting of the alleles present on the first two haplotypes at that site. The IBD sharing pattern was sampled independently for each SNP and thus LD and biological variation in IBD was not modelled. We concatenated the data from many independent simulations to achieve enough data so the IBD sharing was approximately equal to the expected values. Simulation code is available in the supplemental materials.

2.3 | Application to real data sets

To assess the utility of the new method on more realistic data, we applied it to two different publicly available data sets: SNP array data from seven HGDP populations (Rosenberg, 2006), and sequencing data from five related individuals from the Luhya in Webuye, Kenya (LWK) population of the 1,000 Genomes Project phase 3 (The Genomes Project 2015).

2.3.1 | HGDP SNP array

The HGDP SNP array data set was accessed 13 January 2017, and we followed the quality control steps described in Rosenberg (2006) to exclude mislabelled and duplicate samples. We selected seven populations from the HGDP based on the presence of several close familial relationships; five non-African populations: Surui, Pima, Karitiana, Maya and Melanesian, and two African populations: Mbuti Pygmies and Biaka Pygmies.

To ensure a fair comparison to the allele frequency-based inference method in PLINK, and by proxy to other commonly used methods, we constructed data sets where these methods have been shown to perform well. Specifically, we excluded individuals showing obvious signs of admixture (n = 16) or inbreeding (n = 2) from the selected HGDP populations. For details, see Supplemental text 2.3.1. This left us with a total of 142 individuals from the seven populations: Surui (n = 20), Pima (n = 20), Karitiana (n = 21), Maya (n = 16), Melanesian (n = 19), Biaka Pygmies (n = 31) and Mbuti Pygmies (n = 15). For each of these seven populations, we constructed a final set of genotypes by retaining genotypes from autosomal loci with genotyping rate >99%, minor allele frequency (MAF) >5%, Hardy-Weinberg equilibrium p-value > 10^{-4}.

The Ro, R1 and KING-robust kinship statistics for each of the 2,902 within-population pairs of individuals were then calculated from all sites where both individuals had non-missing genotypes.

2.3.2 | 1,000 Genomes sequencing data

To get sequencing data from several different relationship categories, we selected five individuals from two families in the Phase 3 1,000 Genomes (1000G) Luhya in Webuye, Kenya (LWK) population: NA19027, NA19042, NA19313, NA19331, NA19334. Across the five individuals, there is one pair of half-siblings (NA19027 & NA19042), and a separate trio of related individuals with a pair of full-siblings (NA19331 & NA19334), one parent–offspring relationship (NA19313 & NA19331) and another unspecified second-degree relationship (NA19313 & NA19334), possibly avuncular (The Genomes Project 2015). These stated relationships leave six unrelated pairs among the five individuals.

For each pair of the five LWK individuals, we estimated the R0, R1 and KING-robust kinship statistics in five different ways: (1) and (2) by applying the two different sequencing-based approaches described above to the 1000G aligned sequence data files (~4× coverage bam files), (3) by simple genotype counting based on the phased and high-quality curated genotypes provided in the hg37 1000G VCF files, (4) by genotype counting based on the subset of sites in approach 3 that overlap with the Illumina 650Y sites for the HGDP data (to investigate ascertainment, see below) and (5) by calling genotypes from the same 1000G bam files in a basic manner meant to mimic data from a species with a reference genome but few other genetic resources and then simply counting from the called genotypes. For genotype calling, we used samtools mpileup (v1.3.1) to summarize the reads overlapping each position, and bcftools call (v1.3.1) to assign the most likely genotype at each position. We used mostly default settings; non-default flags to samtools specified skipping indel positions. Non-default flags to bcftools specified using the consensus caller. For all sequence-based analyses (1, 2 and 5), we only considered reads with a minimum phred-scaled quality score of 30 and bases with minimum phred-scaled quality score of 20 and we restricted our analyses to genomic regions with a GEM 75mer mappability of 1 (Derrien et al., 2012). Notably, all the methods and filters used here can be applied to any study, including studies with only small contigs, for example made up of RAD loci, making the results relevant beyond resequencing studies utilizing well-assembled genomes.

2.4 | Assessing the effect of SNP ascertainment

To evaluate the effect of SNP ascertainment using real data, we created a subset of the curated genotype data from the five 1,000 Genomes individuals. We selected the sites that overlap with the Illumina 650Y array that was used for the HGDP and estimated our three relatedness statistics. We compared the results from this subset of HGDP sites to

results for the full genotype data set and also to the sequence-based analyses. For an additional comparison, we also performed the same comparison for the methods presented in Rosenberg (2006) and Manichaikul et al. (2010) by constructing scatterplots by their methods.

We also investigated the effect of SNP ascertainment using the data simulated from a constant demography (see "Application to simulated data" for details about the simulations) and compared results for the full data set to results obtained by including only sites with a minor allele count >2 out of 40 chromosomes (MAF > 5%) in the analyses.

2.5 | Assessing the effect of a limited number of sites

To assess the usefulness of the new method on data sets with fewer genomic sites covered by sequencing reads, we constructed reduced size data sets, in a way that mimicked some aspects of reduced-representation sequencing approaches such as RADseq. To produce each reduced data set, we selected a specific number of 200-bp windows randomly from the mappable genomic regions and restricted our analysis to sites falling within them. We used 10 k 50 k, 100 k and 250 k windows, representing ~4× sequencing coverage on 2 M, 10 M, 20 M or 50 M sites, respectively. All other aspects of the analyses were the same, expect for that for these data sets, we applied the IBS- and SFS-approaches to the complete data set, rather than splitting by chromosome as we did for the full data set. We suggest analysing the complete data in single run if your computational resources allow it, as we noticed some upward bias in the estimated number of IBS0 sites when the smaller data sets were analysed separately by chromosome.

2.6 | Comparison to other methods

To get a categorization of relationships for the HGDP data set described above based on a standard, commonly used allele frequency-based method, we first applied the allele frequency-based relatedness estimation algorithm in PLINK (v1.9) (Chang, Chow, Tellier, Vattikuti, & Purcell, 2015) to the individuals from each population separately to estimate the genomewide IBD fractions k_0, k_1 and k_2. Next, we applied the relationship criteria proposed in table 1 of Manichaikul et al. (2010) to the obtained estimates: the estimated k values were combined into an estimate of the kinship coefficient $\theta = \frac{k_1}{4} + \frac{k_2}{2}$, and a relationship degree was assigned to each pair of individuals based on comparing the estimated kinship coefficient to the criteria in the table. Parent–offspring and full-siblings were differentiated based on k_2 values. This provided us with a categorization into five categories: PO, FS, HS, C1 and UR. To achieve additional resolution, we further divided the last category (UR) into two: unknown/distantly related (UK-DR) and unrelated (UN). We did this by simply extending the logic behind the criteria proposed above. Specifically, we set the kinship threshold between UK-DR and UR to $1/2^{13/2}$, which corresponds to including 4th- to 5th-degree relatives in the UK-DR category.

To assess the accuracy and precision of the new method for familial relationship classification within the HGDP data, we examined concordance with the PLINK-based relationship categorization described above. For this purpose, we assigned a relationship category to each pair of individuals in two ways: (1) using the statistics R0 and R1 and (2) using a combination of KING-robust kinship and R0. For the former, we characterized each possible relationship by a single [R1, R0] point generated from data simulated under a demography with a constant population size over time, detailed in the "Application to simulated data" section and assigned each pair of individuals the relationship of the closest point using a Euclidean distance measure. For the latter, we used the KING-robust kinship criteria from table 1 of Manichaikul et al. (2010) as above. Since this table has overlapping kinship ranges for the PO and FS categories, we used the R0 statistic to distinguish PO from FS relationships: Ignoring rare effects like germline mutations and genotyping errors the expected value for R0 for PO relatives is zero, while for FS the value is above 0, we used an ad hoc cut-off of 0.02.

To estimate the statistics for identifying related individuals proposed by Rosenberg (Rosenberg, 2006) and KING (Manichaikul et al., 2010), we note that the KING-robust kinship estimator can (as previously described) be calculated directly from the same nine counts, A–I and so can the fraction of sites IBS0 and IBS2:

$$\text{KING}-\text{robust kinship} = (E - 2(C+G))/(B+D+H+F+2E)$$

$$\text{Fraction IBS0} = (C+G)/(A+B+C+D+E+F+G+H+I)$$

$$\text{Fraction IBS2} = (A+E+I)/(A+B+C+D+E+F+G+H+I)$$

We used these formulas in all our comparisons because this allowed us to estimate these statistics not only from genotype data but also directly from sequence data in the same manner as for R0 and R1. However, we note that this is our approach to estimating those statistics, and that existing tools like KING only allow users to estimate the statistics from genotype data.

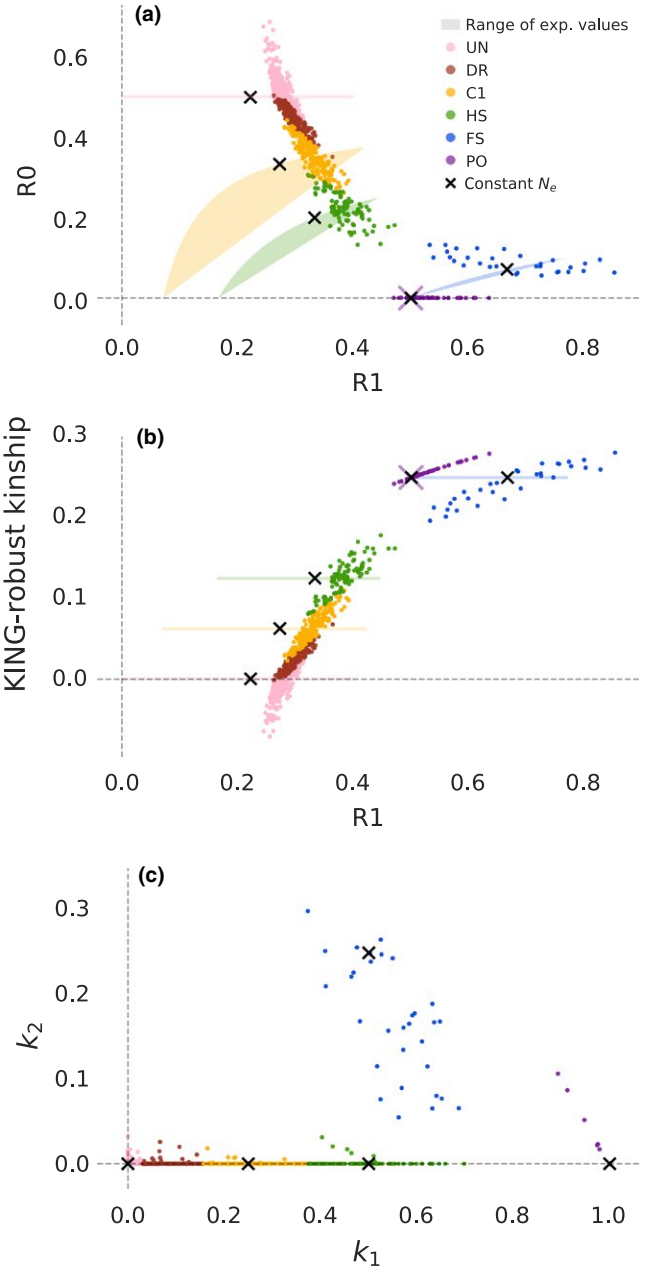
3 | RESULTS

To assess the performance of the new method, we first applied it to simulated genotype data to ensure that it works on sufficient data and to assess how sensitive it is to the underlying demographic history of the population the analysed samples are from. Next, we applied the new method to real data from different platforms to assess its performance on more realistic data. Finally, we performed a couple of additional analyses to access how robust the new method is to SNP ascertainment and to having data from only a limited number of sites available. Below we describe the results of all these analyses.

3.1 | Application to simulated data

We first applied the new method to simulated genotype data from several different relationship pairs from populations with three

different demographic histories: (1) constant N_e, (2) 10-fold increase in N_e over the past 100 generations, and (3) 10-fold decrease in N_e over the past 100 generations. We see very similar results across all three demographic scenarios, and in all cases, the R1–R0 and R1-KING-robust kinship values obtained from the simulated data were within the theoretically derived ranges of expectations (Figure 2). These results demonstrate that the method works if sufficient high-quality genotype data are available. Furthermore, they demonstrate that the range of examined population size histories has a limited effect, even though demographic history affects the allele frequency spectrum. In turn, this suggests that the range of expected values that are realistic for real data is markedly smaller than the theoretically possible ranges also shown in the figure, which is useful for classification purposes.



3.2 | Application to SNP array data

Next, we applied the method to SNP array data from the Human Genome Diversity Project (HGDP) to see how the method works on real data, for which standard allele frequency-based methods like PLINK are known to perform well. More specifically, we applied it to genotype data from unadmixed and non-inbred samples from seven populations originating from the HGDP. This resulted in the R1–R0 and R1-KING-robust kinship plots shown in Figure 3a,b (for population-specific plots, see Supplementary Figures S1 and S2). The true relationships for the pairs of individuals are not known; we instead coloured each point in Figure 3a,b according to the relationship category inferred based on results from the standard, commonly used allele frequency-based method PLINK (Figure 3c, Supplementary Figure S3). Since there are least 15 individuals in each of the seven selected HGDP populations, the allele frequency estimates for these populations should be reasonably accurate even with some relatedness. Hence, with the large amount of data available in this data set, the allele frequency-based method should provide correct inference of most, if not all, pairs closer than first cousins, but may not be able to fully distinguish first cousins from more distantly related pairs.

In Figure 3a,b, points from each relationship category clearly cluster together on both the R1–R0 plot and the R1-KING-robust kinship plot. Moreover, these clusters are located near both their theoretically derived ranges of expected values and the values from simulated data in a similar manner, the k_1–k_2 values for the same pairs of individuals cluster close to the expected and simulated values of k_1 and k_2 (Figure 3c). Almost all pairs identified as parent–offspring (PO) by the frequency-based method are easy to identify as such in both the R1–R0 plot and the R1-KING-robust kinship plot, which is not the case when only a single statistic is used (see also Supplemental Figures S1 and S2). The same is true for full-siblings (FS). Furthermore, points classified as half-siblings/avuncular/grandparent–grandchild (HS) or first cousins (C1) by the frequency-based method have a minimal overlap with each other and with less-related pairs (Figure 3a,b). The few

FIGURE 3 R1–R0 and R1-KING-robust kinship scatterplots for seven HGDP populations. Each coloured point represents a pair of individuals and is coloured according to the relationship category inferred using an allele frequency-based approach. Coloured shaded areas/lines show the theoretically derived range of expected values for specific relationship categories, as in Figure 2. Black "X"s show the values for a pair of individuals simulated under a constant population size, as in Figure 2. Note that in addition to the relationship categories for Figure 2 there is an additional category here representing distantly related pairs (DR). (a) R1–R0 plot for all pairs of individuals within each population (b) R1-KING-robust kinship plot for all pairs of individuals within each population. (c) Scatterplot of the two relatedness coefficients k_1 and k_2 for all pairs of individuals within each population estimated using the allele frequency-based approach implemented in PLINK. Note that the black "X"s here show simulated values for k_1 and k_2 and are not inferred by PLINK, they approximately coincide with the expected values of k_1 and k_2 for each relationship category (Supplementary Table S1)

pairs of individuals that were difficult to classify are the same pairs as those that are edge cases for the allele frequency-based method. This is apparent in an R1–R0 plot of the HGDP data constructed excluding pairs that are closer than 0.01 to the kinship coefficient thresholds that the frequency-based method used when classifying relationships (Supplementary Figure S4).

To quantify precision and accuracy, we examined the concordance between classifications based on the new method and the PLINK-based classification. We tried two simple classification schemes: one based on R1–R0, which uses proximity to the values we obtained from simulated data from a constant N_e demography, and one based on KING-robust kinship (for details see Methods and Materials). The results supported the visual assessment: both classification schemes are highly concordant with the classifications obtained using the frequency-based method (Supplemental Figure S5). Mean precision across all relationship categories was 0.90 for the R1–R0 method, vs. 0.89 for KING-robust kinship. Mean recall across all relationship categories was 0.88 for R1–R0, vs. 0.89 for KING-robust kinship. The relationship categories for which the method has the lowest precision are the first cousins vs. less-related pairs, where the allele frequency-based method is also known to have a hard time making classifications. For PO, FS and UR alone, the mean precision is as high as 0.99 for R1–R0 and 0.96 for KING-robust-kinship, and the mean recall for these three categories is as high as 0.96 for R1–R0 and 0.99 for KING-robust kinship. Hence, the new method provides comparable performance to a frequency-based method when sufficient genotype data are available, but without the need for allele frequency information.

3.3 | Application to sequencing data

To assess how well the new method works on more limited real data, we applied it to sequencing data from five low-depth (~4×) human genomes from the 1,000 Genomes project. Among the five selected samples, there is a parent–offspring pair, a pair of full-siblings, a pair of half-siblings, an unspecified 2nd-degree relationship (e.g., avuncular), and the rest are unrelated. We estimated the R0, R1 and KING-robust kinship for each pair in several ways. First, by using an IBS-based approach that estimates the proportion all pairwise combinations of the 10 possible genotypes (Figure 4, "IBS"). Second, by using an SFS-based approach where we estimated the two-dimensional site-frequency spectrum (2D-SFS) of each pair with a bi-allelic model and calculated R0, KING-robust kinship and R1 based on this spectrum (Figure 4, "realSFS"). Both these approaches base their estimates on genotype likelihoods calculated from the sequencing read data, instead of called genotypes, and take the uncertainty of the underlying genotypes that is inherent to low-depth sequencing data into account. The key difference between them is that the SFS-based approach requires specification of an allele known to exist at each site, whereas the IBS-based approach has no such requirement, making it more generally applicable. The approaches also differ in how they deal with sites with more than two unique alleles, either excluding them (IBS-based approach) or integrating over the two-allele possibilities (SFS-based approach), but these sites are rare (mean fraction as estimated by IBS: 1.8E-6) so the impact of discarding them is minimal.

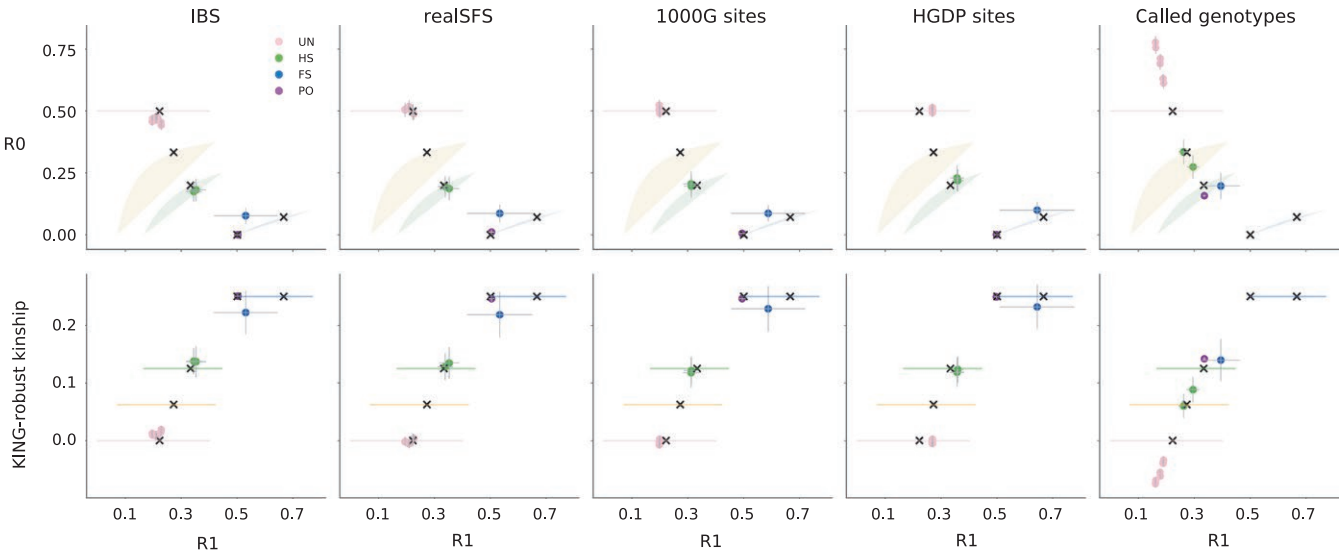


FIGURE 4 Relatedness plots for all pairs among five LWK individuals from the 1000G Project. (Top) R1–R0 scatterplots for pairs of five LWK individuals for five different analysis approaches: (1) IBS: estimation from (~4×) 1000G bam files, (2) realSFS: site-frequency spectrum-based estimation from (~4×) 1000G bam files, (3) 1000G sites: genotype counting using curated 1000G genotypes from the 1000G project, (4) HGDP sites: genotype counting using curated 1000G genotypes but only at sites that overlap with the Illumina 650Y array used for the HGDP, and (5) called genotypes: genotype counting using genotypes called de novo from (~4×) 1000G bam files. Points are coloured by their true relationship status, as reported by 1000G. Thin grey lines show confidence intervals (±2 SE) estimated using a chromosome jackknife. Coloured shaded areas/lines show the theoretically derived range of expected values for specific relationship categories from Figure 2. Black "X"s show the values for pairs with different relationships simulated under a constant population size, as in Figure 2. (Bottom) R1-KING-robust kinship scatter plots for the same data sets, confidence intervals and expected ranges are constructed in the same way

With the values estimated using the SFS-based approach, it is possible to visually classify of all the pairs to their relationship category within the set of close familial relationships (PO, FS, HS or UR), or by using one of the classification methods introduced earlier. Results for the IBS-based approach were similar, but unrelated individuals have a slight decrease in R0 and a slight increase in R1 and KING-robust kinship, compared to the SFS-based approach. This makes unrelated individuals appear slightly more related than expected for unrelated individuals from a homogenous population. However, despite this bias, it is still possible to correctly classify of all the pairs to their relationship category, suggesting that the IBS-based approach can be used when not enough information is available for the SFS-based approach. The chromosome block-jackknife estimates of uncertainty for the genotype likelihood-based methods were small, and varied by relationship type, with the pair of full-siblings having the most uncertainty in R0, R1 and KING-robust kinship.

We also calculated the three statistics from the high-quality phased genotypes for the same five individuals available from the 1,000 Genomes Project Phase 3 (Figure 4, "1000G sites") to see how well the two genotype likelihood-based approaches applied to low-depth sequencing data perform compared to direct calculations from high-quality genotype data for the same samples. In this comparison, results obtained by using the genotype likelihood-based approaches applied to low-depth sequencing data are close to those obtained from the high-quality genotypes for all the pairs (Figure 4).

Finally, we also made R1–R0 and R1-KING-robust kinship plots based on genotypes that we obtained through a standard genotype calling procedure from the raw read data. We did this to investigate whether the genotype likelihood-based approaches are necessary or one could just as well use genotypes called from the ~4× data. As expected, genotype calling had a large negative effect on the outcome; in the resulting R1–R0 and R1-KING-robust kinship plots, the half-siblings appear within the range of expected values for first cousins and both the parent–offspring and full-sibling pairs appear within the range of expected values for half-siblings (Figure 4, "called genotypes"). These results demonstrate the pitfalls of basing any relationship inferences, including R1–R0 and R1-KING-robust kinship plots, on genotypes called from low-depth data. Notably, this is also the case for the methods presented in Rosenberg (2006) and Manichaikul et al. (2010) (Figure 5, "called genotypes"). This clearly demonstrates that, with ~4× sequencing data, calling genotypes without external information, such as an imputation reference panel, is not a good alternative to a genotype likelihood-based approach. This implies that software packages designed to work only on genotype data, such as KING, should not be used on data like this.

3.4 | Assessing the effect of SNP ascertainment

To assess the effect of SNP ascertainment, we applied the new method to three different subsets of data from the five 1,000 Genomes

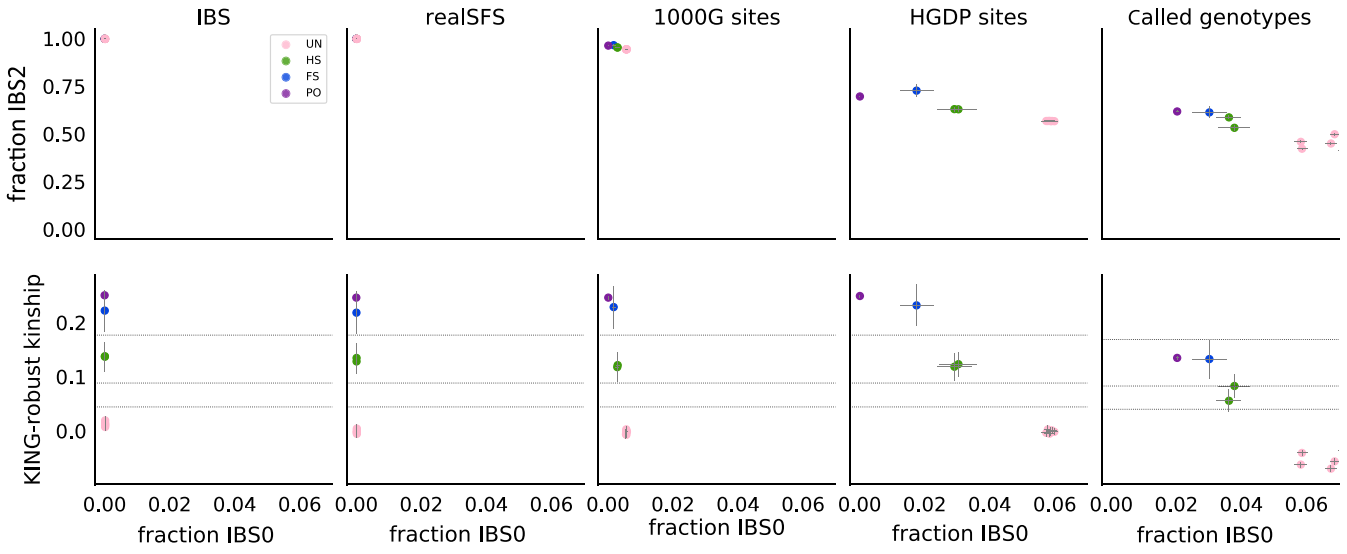


FIGURE 5 Results from two alternate frequency-free methods to different subsets and types of data from five 1000 Genomes samples. (Top) Results from applying the plotting approach from Rosenberg (2006) to pairs of the same five LWK individuals for five different analysis approaches: 1) IBS: estimation from (~4×) 1000G bam files, 2) realSFS: site-frequency spectrum based estimation from (~4×) 1000G bam files, 3) 1000G sites: genotype counting using curated 1000G genotypes from the 1000G project, 4) HGDP sites: genotype counting using curated 1000G genotypes at sites that overlap with the Illumina 650Y array used for the HGDP and 5) genotype counting using called genotypes: genotype called de novo from (~4×) 1000G bam files. Pairs are coloured by their true relationship status, as in Figure 3. Fraction IBS0/IBS2 are the overall fraction of sites that are IBS0/IBS2, res pectively. Grey lines centred on each point show confidence intervals (±2 SE) based on a chromosome jackknife. (Bottom) Results from applying the KING-robust based approach to the same pairs of LWK individuals using the same five different analysis methods as above. The horizontal black lines show the kinship thresholds used to distinguish unrelated (UR), first cousins (C1) half-siblings (HS), full-siblings (FS) and (PO) following (Manichaikul et al., 2010) from bottom to top, respectively. Thin grey lines centred on each point show confidence intervals (±2 SE) estimated using chromosome jackknife

individuals. The results for each of the three ascertainment schemes; all sites covered by sequencing data, 1,000 Genomes release sites and Illumina 650Y SNP array sites are similar (left four panels, Figure 4), showing SNP ascertainment does not have a large effect.

For comparison, we performed the same assessment of the methods presented in Rosenberg (2006) and Manichaikul et al. (2010) by constructing scatterplots of the same type as those shown in their papers (Figure 5). This revealed that both these other methods are much more affected by ascertainment than the method proposed here. In particular, the Rosenberg method is affected on both its *x*-axis (IBS0) and its *y*-axis (IBS2), which means that the expected region of the plot for each relationship will be different for different data sets (top row of Figure 5). The method presented in Manichaikul et al., 2010 is affected by the SNP ascertainment mainly on its *x*-axis (IBS0, bottom row of Figure 5). Therefore, the ascertainment mainly affects the ability to distinguish between parent–offspring and full-siblings, since the *y*-axis, which is only slightly affected by ascertainment, is the kinship coefficient, which can be used to distinguish between most close relationships except for parent–offspring and full-siblings. The *x*-axis, IBS0, is included in part to help make the distinction between PO and FS (Manichaikul et al., 2010), but this ability is clearly affected by SNP ascertainment (bottom row of Figure 5).

To further explore the effect of SNP ascertainment on the new method, we also performed analyses of the previously mentioned simulated of data from a population with a constant population size. This time we only analysed SNPs with MAF above 5% and compared the results to the results for the full data set. This confirmed the results from the real data analyses: SNP ascertainment does change the values a bit compared to when all sites are analysed, however the change is limited (Figure 2). This is well in line with the fact that we got very similar results for the simulated data from three populations with quite different population size histories and consequently different allele frequency spectra. Indeed, the effect of population size decline is similar to that of ascertaining for common SNPs, which makes sense because population decline is known to lead to a skew in the allele frequency spectrum towards more common SNPs.

3.5 | Assessing the effect of a limited number of sites

Genomewide shotgun sequencing data, as is available for the 1000G individuals, is not available for all species. Studies may instead have RADseq or similar data, covering only a fraction of genomic sites. To assess to what extent the new method can be used to analyse such data sets, we performed analyses of subsets of the 1000G data, constructed to mimic RAD sequencing data. Specifically, we analysed four subsets that consisted of 10 k, 50 k, 100 k and 250 k, 200 bp windows, representing 2 M, 10 M, 20 M or 50 M sites, respectively. For all but the smallest data subset, the point estimates were similar to those obtained using the full data set, showing the method is applicable when reducing the number of sites even with ~4× coverage (Figure 6, supplemental file 1). This suggests that even with the reduced number of sites tested, there was sufficient data to characterize the genomewide mean IBD fractions for both closely related and unrelated pairs. The uncertainty in the estimates, as estimated by a chromosome jackknife, increased with fewer sites, but the effect was limited, suggesting the biological variation in IBD sharing across chromosomes was larger than sampling variance across the examined sites.

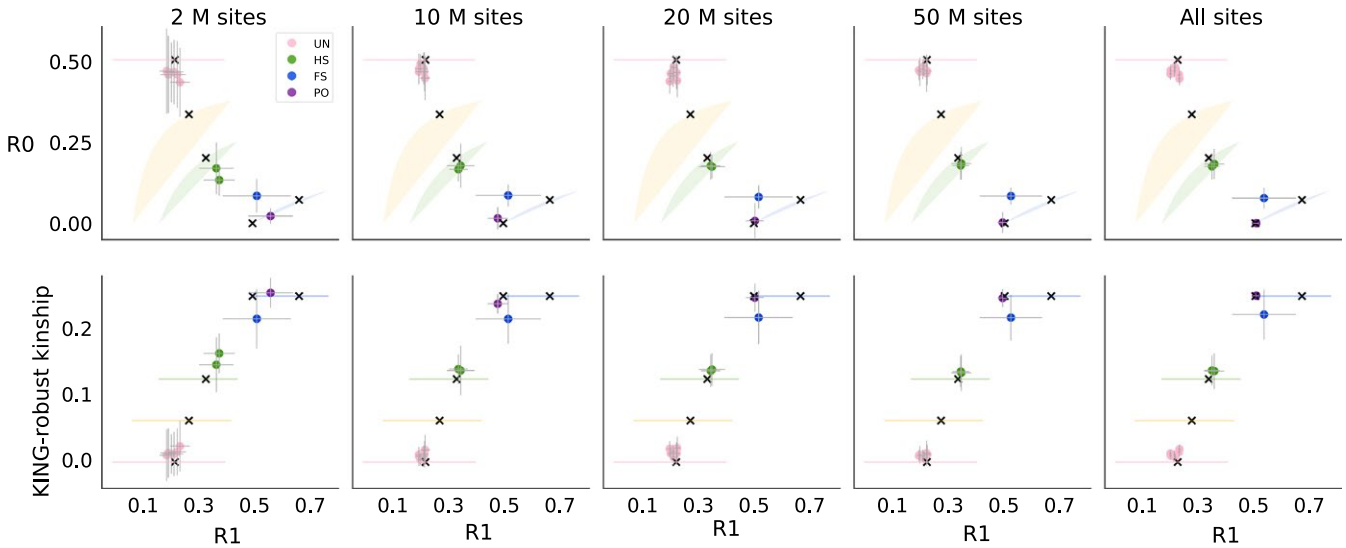


FIGURE 6 The effect on estimates of R0, R1 and KING-robust kinship from reducing the number of sites covered by sequencing data from the same five LWK individuals as in Figures 4 and 5. Each point shows the point estimate and error bars show ±2 SE estimated using chromosome jackknife. Each column shows results for different numbers of examined basepairs, including non-variable sites. Pairs of individuals are coloured by their true relationship. These plots show the results of the IBS-based method, see supplemental for results from the SFS-based approach. Coloured shaded areas/lines show the theoretically possible range of expected values for specific relationship categories from Figure 2 and black "X"s show the values for different relationship pairs of individuals simulated under a constant population size, as in Figure 2

4 | DISCUSSION

We have presented a simple new method for inferring if and how two individuals are related based solely on genetic data from the two target individuals. We demonstrated two ways in which it can also be applied directly to sequencing data via genotype likelihoods. And importantly, we showed that, the method provides useful results when applied to ~4× sequencing data as well as RADseq like subsets of such data. All of this combined implies that—unlike previous methods—this new method can be used even if all you have is low-depth sequencing data from a few individuals from a species without a reference genome.

4.1 | Comparison to similar methods

The new method is based on plotting two statistics, R0 and KING-robust kinship against a third, new statistic R1. The R0 statistic similar to the test statistic proposed by Lee (2003) to test for relatedness. The only differences are that the numerator and denominator are flipped and that E, the proportion of sites where both individuals are heterozygous, is included in the denominator in the statistic defined by Lee but absent in R0. R0 is also similar to the pairwise population concordance (PPC) statistic in PLINK (Purcell et al., 2007), a test of if the genotypes of a pair of individuals have more IBS0 sites than two unrelated individuals with the same ancestry are expected to have, signalling they have different ancestries. The KING-robust kinship estimator was proposed by Manichaikul et al. (2010) and implemented for genotype data in the program KING. Here, we extend it to estimation directly from sequencing data. Notably, our results suggest that this extension is vital for successful application to low-depth sequencing data, because estimates based on genotypes called from low-depth sequencing data are very poor (rightmost panels of Figures 4 and 5), which makes programs like KING inappropriate to apply to such data. This extension can also be used for similar pairwise statistics and thus makes existing methods based on such statistics, like KING-robust kinship, more widely applicable.

However, this extension is not the only contribution of this study. Another key new contribution is to provide an alternative to the IBS0 statistic (the proportion of sites where the two individuals share zero alleles IBS) that was utilized by Rosenberg (2006) and Manichaikul et al. (2010). As we have shown, the fraction of sites that are IBS0 or IBS2 is very sensitive SNP ascertainment, meaning that results are only comparable within each ascertainment scheme. The method from Rosenberg (2006), where IBS0 is combined with IBS2, is difficult, if not impossible, to use for relatedness inference in general because the fraction of sites that are IBS2 and IBS0 varies so wildly across different ascertainment schemes, such as between SNP arrays and sequencing data. On the other hand, KING-robust kinship is still very useful, but it loses the ability to distinguish between parent–offspring and full-siblings, as IBS0 was used for this. Due to this sensitivity to ascertainment, samples cannot be analysed in isolation and must be placed in the context of other samples with known relationships and the same ascertainment scheme. This requirement makes it difficult to apply these previous methods to ancient humans or other species with limited sample sizes.

In contrast, the ability to identify relatives based on expected values is maintained in the new method, regardless of ascertainment scheme due to the use of R0, instead of IBS0, which makes the new method robust to SNP ascertainment. Parent–offspring pairs tend to have an R0 estimate extremely close to 0, making them particularly easy to identify via the R1–R0 plot. The R1-KING-robust kinship plot, on the other hand, has the appealing aspect that the kinship axis has a biological interpretation, defined as the probability that two alleles sampled at random from two individuals are identical by descent. Hence, the two plots types, R1–R0 and R1-KING-robust kinship, each have their advantages. Finally, it is worth noticing that the two plots types seem to work better than a range of other plots constructed from similar ratio statistics that we explored (Supplementary Figure S6).

4.2 | Limitations and applications

While the new method provides substantial advantages over previous methods in situations with limited data, it does have some limitations. First, like most other relatedness inference methods, such as PLINK, the proposed method assumes that the individuals are not inbred and that they originate from the same homogenous population. And like many other relationship inference methods, it is not necessarily robust to violations of these assumptions. Previous studies have shown the effect of population structure and admixture on relatedness inference is complex and can potentially lead to bias in either direction depending on the circumstances, and this is true even for KING, which was developed to be robust to population structure (Conomos et al., 2016; Ramstetter et al., 2017; Thornton et al., 2012). Specific methods have been developed to correct for admixture when the allele frequencies in the admixing populations are known (e.g., Thornton et al., 2012; Moltke & Albrechtsen, 2014), or enough samples are available (Conomos et al., 2016; Dou et al., 2017). But since these methods work by exploiting knowledge about allele frequencies or access to many samples for their correction, the pairwise R0, R1 and KING-robust kinship statistics cannot be easily corrected in a similar manner. However, we note that Lee (2003) showed that the statistic he proposed for testing for relatedness can also be used to detect if two unrelated samples are not from the same homogeneous population. If this is the case, Lee's statistic will be significantly smaller than 2/3; and equivalently R0 will be significantly above 0.5, which may be helpful when interpreting R0, R1, KING-kinship plots in the presence of admixture or population structure more generally. Regarding inbreeding, one potential way to assess if one of the individuals is inbred is to compare heterozygosities across individuals; non-inbred and non-admixed individuals from the same population should have similar heterozygosity, so marked heterozygosity differences can be a warning signal.

A second limitation, which is shared with other relatedness estimation methods, is that there is significant biological variation in the

amount of IBD sharing between relatives with the same pedigree relationship due to randomness inherent in the process of recombination (Hill, 1993; Rasmuson, 1993). For humans, this means that a pair of relatives, say first cousins, will sometimes share less of their genomes IBD than another pair with a more distant pedigree relationship, say second cousins. This makes classification into specific relationships difficult. The degree of biological variation in IBD sharing between relatives varies across species and can even differ between sexes due to sex-specific recombination patterns. This makes it difficult to provide general guidance appropriate for all species. In general, species with more chromosomes and more recombination will have less variation in IBD sharing for a defined pedigree relationship, making it easier to distinguish among various potential relationship categories. To quantify this uncertainty, we propose a chromosomal bootstrap procedure that can be used if reads can be assigned to chromosomes.

Biological variation in IBD sharing is also related to the estimation and interpretation of confidence intervals on statistics like R0, R1 and KING-robust kinship. Relatedness and limited recombination also cause correlation between sites in the genome, due to shared IBD segments and LD. This correlation between sites increases the variance in the estimates of these statistics in a way that can be difficult to fully account for when computing confidence intervals. For statistics that test for introgression such as the D-statistic (Patterson et al. 2012), where the main concern is correlation due to LD, a block jackknife, leaving out contiguous blocks (e.g., 5 Mb) is a common approach. When considering relatedness, we want to compare our estimates to the expectations of each relationship category. Since shared IBD segments can be much longer than the range of LD we propose a more appropriate chromosome jackknife. In either case, a jackknife (or bootstrap) over single sites will fail to provide a confidence interval that accounts for the non-independence of the sites. For more discussion on this topic, see Thompson (2013). Unfortunately, this means that it is difficult to provide the most appropriate confidence intervals when no information about genomic positions is available.

Despite these limitations, we believe that the results presented here suggest the new method constitutes a helpful new tool for relatedness inference for studies with limited data. Identifying related samples is a crucial step in nearly any genetic analysis and can also reveal other problems such as duplicate samples or cross-contamination of genetic material. Removing the requirements to specify allele frequencies and to have accurate genotypes has the potential allow the identification of relatives even in small studies of non-model species or ancient samples. These types of studies do not currently have many good options to address relatedness.

AUTHOR CONTRIBUTIONS

AA, IM and RKW designed the research. RKW performed research, with input from IM and AA. IM and RKM wrote the paper together, with help from AA.

DATA ACCESSIBILITY

The IBS method is available at: http://www.popgen.dk/software/index.php/IBSrelate.

The data sets used are publicly available.

The HGDP SNP array data are available at: ftp.cephb.fr/hgdp_supp1.

The 1000G phase 3 aligned sequencing data are available at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data.

The 1000G phase 3 called genotypes are available at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

ORCID

Ryan K. Waples https://orcid.org/0000-0003-0526-6425
Anders Albrechtsen https://orcid.org/0000-0001-7306-031X
Ida Moltke https://orcid.org/0000-0001-7052-8554

REFERENCES

Abecasis, G. R., Cherny, S. S., Cookson, W., & Cardon, L. R. (2001). GRR: Graphical representation of relationship errors. *Bioinformatics*, *17*, 742–743. https://doi.org/10.1093/bioinformatics/17.8.742

Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F. C., & Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology*, *33*, 266–274. https://doi.org/10.1002/gepi.20378

Baca, M., Doan, K., Sobczyk, M., Stankovic, A., & Weglenski, P. (2012). Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community. *BMC Genetics*, *13*, 30. https://doi.org/10.1186/1471-2156-13-30

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*, 781–791. https://doi.org/10.1038/nrg1916

Busing, F. M. T. A., Meijer, E., & Van Der Leeden, R. (1999). Delete-m jackknife for unequal m. *Statistics and Computing*, *9*, 3–8.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics*, *98*, 127–148. https://doi.org/10.1016/j.ajhg.2015.11.022

Derrien, T., Estelle, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome mappability. *PLoS ONE*, *7*, e30377. https://doi.org/10.1371/journal.pone.0030377

Dou, J., Sun, B., Sim, X., Hughes, J. D., Reilly, D. F., Tai, E. S. ... Wang, C. (2017). Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS Genetics*, *13*, e1007021. https://doi.org/10.1371/journal.pgen.1007021

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L. ... Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, *19*, 318–326.

Hill, W. G. (1993). Variation in genetic identity within kinships. *Heredity*, *71*, 652–653. https://doi.org/10.1038/hdy.1993.190

Hill, W. G., & Weir, B. S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, *93*, 47–64. https://doi.org/10.1017/S0016672310000480

Kardos, M., Luikart, G., & Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity*, *115*, 63–72. https://doi.org/10.1038/hdy.2015.17

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, *12*, e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356. https://doi.org/10.1186/s12859-014-0356-4

Korneliussen, T. S., & Moltke, I. (2015). NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, *31*, 4009–4011. https://doi.org/10.1093/bioinformatics/btv509

Kuhn, J. M. M., Jakobsson, M., & Gunther, T. (2018). Estimating genetic kin relationships in prehistoric populations. *PLoS ONE*, *13*, e0195491.

Lee, W. C. (2003). Testing the genetic relation between two individuals using a panel of frequency-unknown single nucleotide polymorphisms. *Annals of Human Genetics*, *67*, 618–619. https://doi.org/10.1046/j.1529-8817.2003.00063.x

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110

Moltke, I., & Albrechtsen, A. (2014). RelateAdmix: A software tool for estimating relatedness between admixed individuals. *Bioinformatics*, *30*, 1027–1028. https://doi.org/10.1093/bioinformatics/btt652

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, *7*, e37558. https://doi.org/10.1371/journal.pone.0037558

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575. https://doi.org/10.1086/519795

Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J. … Williams, A. L. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, *207*, 75–82.

Rasmuson, M. (1993). Variation in genetic identity within kinships. *Heredity*, *70*, 266–268. https://doi.org/10.1038/hdy.1993.38

Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, *70*, 841–847. https://doi.org/10.1111/j.1469-1809.2006.00285.x

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A. … Willerslev, E. (2017). Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science*, *358*, 659–662.

Stevens, E. L., Heckenberg, G., Roberson, E. D., Baugher, J. D., Downey, T. J., & Pevsner, J. (2011). Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genetics*, *7*, e1002287. https://doi.org/10.1371/journal.pgen.1002287

The Genomes Project. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of Human Genetics*, *39*, 173–188. https://doi.org/10.1111/j.1469-1809.1975.tb00120.x

Thompson, E. A. (2000). *Statistical inferences from genetic data on pedigrees NSF-CBMS regional conference series in probability and statistics (Vol. 6)*. Beachwood, OH: IMS.

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, *194*, 301–326. https://doi.org/10.1534/genetics.112.148825

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. *American Journal of Human Genetics*, *91*, 122–138. https://doi.org/10.1016/j.ajhg.2012.05.024

Wiuf, C. (2006). Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology*, *53*, 821–841. https://doi.org/10.1007/s00285-006-0031-0

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Waples RK, Albrechtsen A, Moltke I. Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Mol Ecol*. 2019;28:35–48. https://doi.org/10.1111/mec.14954